

Automated Fake News Detection Using Machine Learning and BERT

Aharon Rabson

Business Problem

The spread of fake news has become a significant challenge in today's digital ecosystem, influencing political outcomes, shaping public opinion, and even impacting health-related decisions. Online platforms accelerate the dissemination of misinformation at a speed and scale that traditional manual fact-checking cannot match. Detecting fake news manually is resource-intensive and impractical for the vast volume of daily content.

This project aims to demonstrate an automated, explainable system for distinguishing fake from real news using both classical machine learning models and modern deep learning approaches. The solution emphasizes not only high predictive performance but also interpretability, ensuring that decision-making processes remain transparent and trustworthy.

Background / History

While fabricated stories have existed throughout history, the rise of digital and social media has amplified their reach and influence. Social platforms enable instantaneous sharing, allowing false information to go viral before fact-checkers can intervene. Traditional detection methods, such as manual review by journalists or specialized organizations, struggle to keep pace with the sheer volume of content.

Machine learning models have emerged as powerful tools in this space, capable of identifying linguistic and structural patterns that differentiate credible reporting from fabricated stories. Classical methods such as Logistic Regression provide speed and interpretability, while advanced architectures like BERT can capture deep contextual nuances in language, offering state-of-the-art accuracy.

Data Explanation

Data Sources

Kaggle Fake and Real News Datasets:

The dataset comes from the **Kaggle Fake and Real News Datasets**, containing approximately 40,000 articles labeled as either "fake" or "real."

Key fields include:

- **Title** – Article headline.
- **Text** – Full article content.
- **Subject** – Category or topic of the news.
- **Date** – Publication date.
- **Label** – Binary indicator (0 = fake, 1 = real).

Data Preparation Steps

To prepare the dataset for modeling, several preprocessing steps were applied:

1. **Data Cleaning** – Removal of null values and duplicates to ensure integrity.
2. **Text Normalization** – Lowercasing, punctuation, and number removal to maintain consistency.
3. **Stopword Filtering** – Elimination of common, non-informative words.
4. **Lemmatization** – Reducing words to their base form to unify similar terms.
5. **Label Encoding** – Assigning numeric labels for machine learning compatibility.

These preprocessing steps ensured the text data was in a standardized format suitable for both classical machine learning and deep learning models.

Data Dictionary

Column	Description
title	Article headline
text	Article body
subject	News topic/category
date	Publication date
label	0 = Fake, 1 = Real
clean_text	Preprocessed full text

Methods

The methodology combined exploratory data analysis (EDA), classical machine learning models, deep learning with BERT, and explainability techniques to produce an end-to-end detection pipeline.

- EDA helped identify patterns in word usage, topic distribution, and label balance.
- Classical Models used TF-IDF vectorization to convert text into numerical features, followed by Logistic Regression and Random Forest for classification.
- Evaluation Metrics included accuracy, ROC-AUC, confusion matrix analysis, and F1-score to provide a balanced performance assessment.
- Deep Learning leveraged BERT fine-tuning via the Hugging Face Transformers library to capture nuanced language patterns.
- Explainability was achieved through SHAP (SHapley Additive exPlanations), allowing token-level interpretation of BERT predictions.

[illegible]

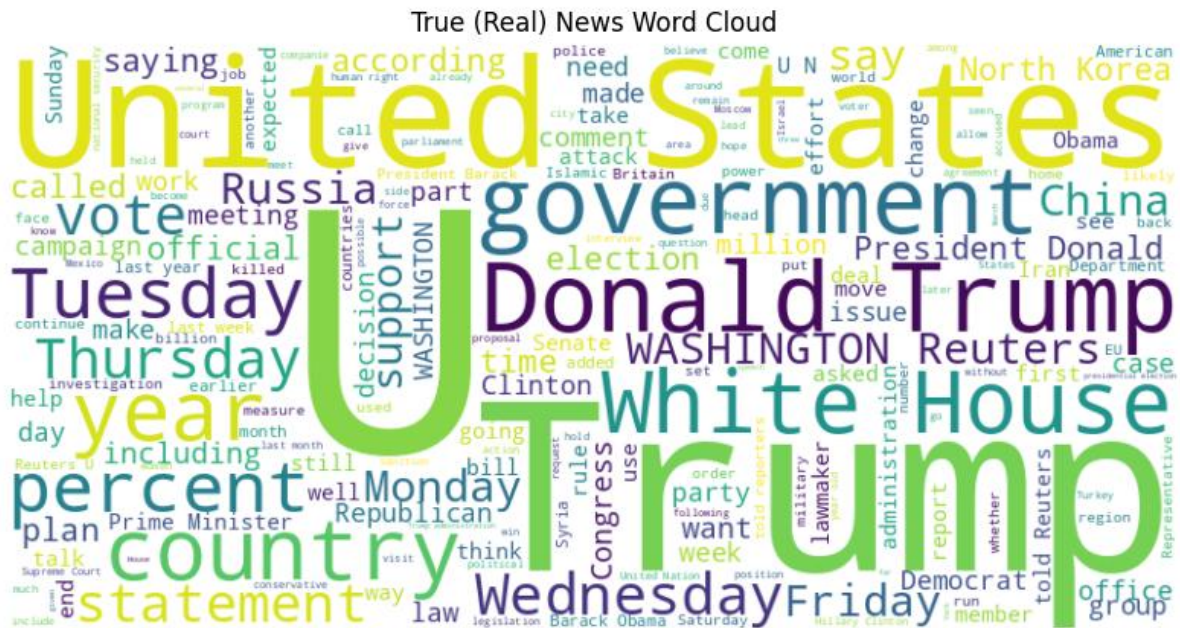


Figure 3. Word cloud for True News (top words after cleaning).

Findings:

The dataset was found to be moderately balanced between fake and real news. Word cloud visualizations revealed that fake news frequently contained emotionally charged or viral-oriented terms such as “*video*,” “*image*,” “*via*,” and “*breaking*.” In contrast, real news headlines were more likely to mention reputable agencies, dates, and formal political references, such as “*Reuters*” and “*Washington*.”

Classical Model Results

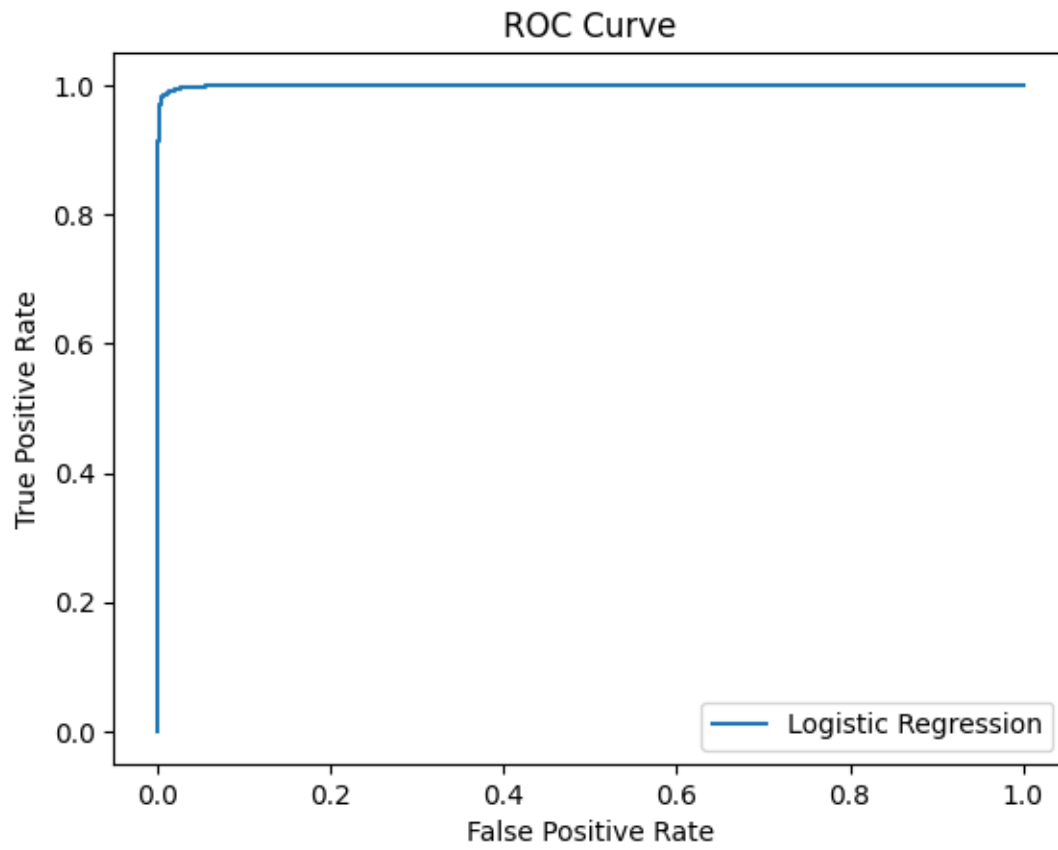


Figure 4. ROC curve for *Logistic Regression* model.

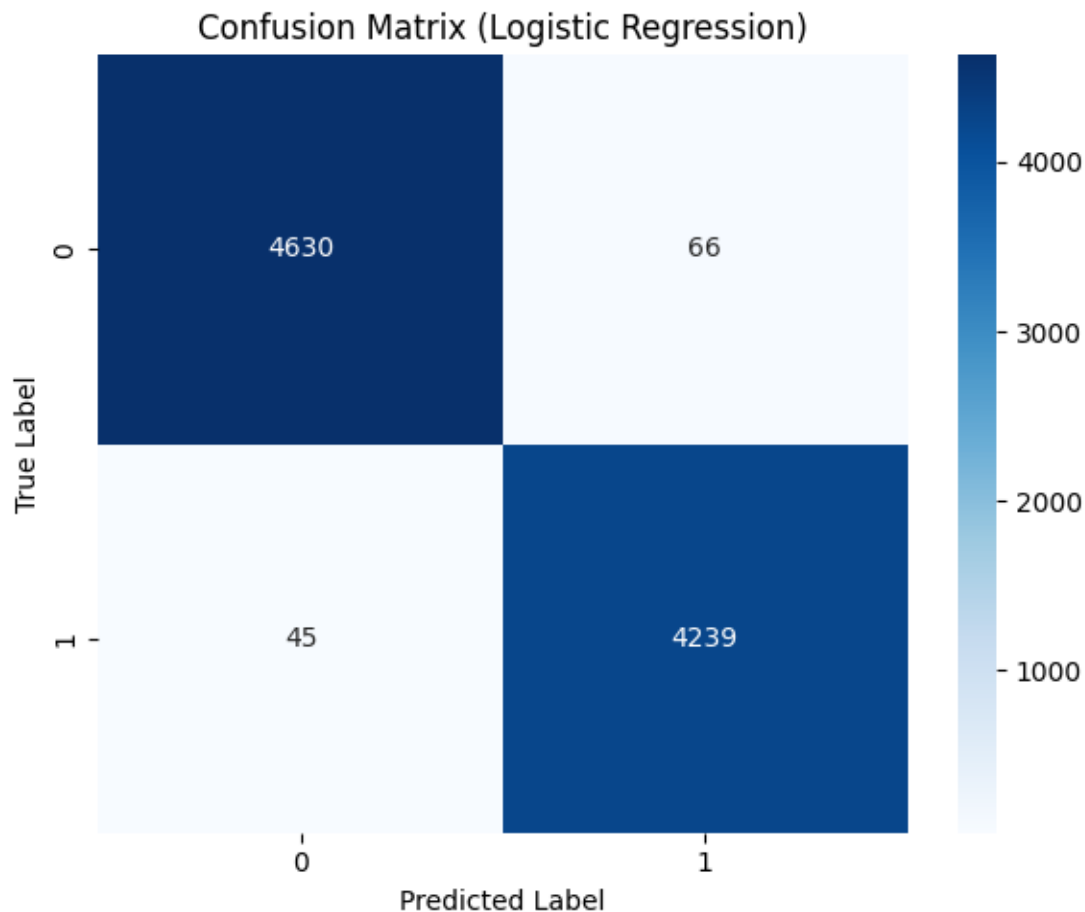


Figure 5. *Confusion Matrix for Logistic Regression.*

Performance:

Logistic Regression achieved approximately **97% accuracy** with an ROC-AUC exceeding **0.97**, demonstrating strong separability between the two classes. Random Forest achieved similar results but with slightly lower interpretability.

Feature importance analysis for Logistic Regression revealed that terms like “*reuters*”, “*wednesday*”, and “*washington*” strongly indicated real news, while words like “*video*”, “*image*”, and “*hillary*” were strong predictors of fake news.

The confusion matrix showed that the model correctly identified the vast majority of both fake and real headlines, with very few false positives and false negatives. This suggests the model is not only accurate but also consistent across both classes.

Top words predicting REAL news:

1. reuters: 23.589
2. washington reuters: 10.014
3. wednesday: 6.176
4. president donald: 5.816
5. tuesday: 5.770

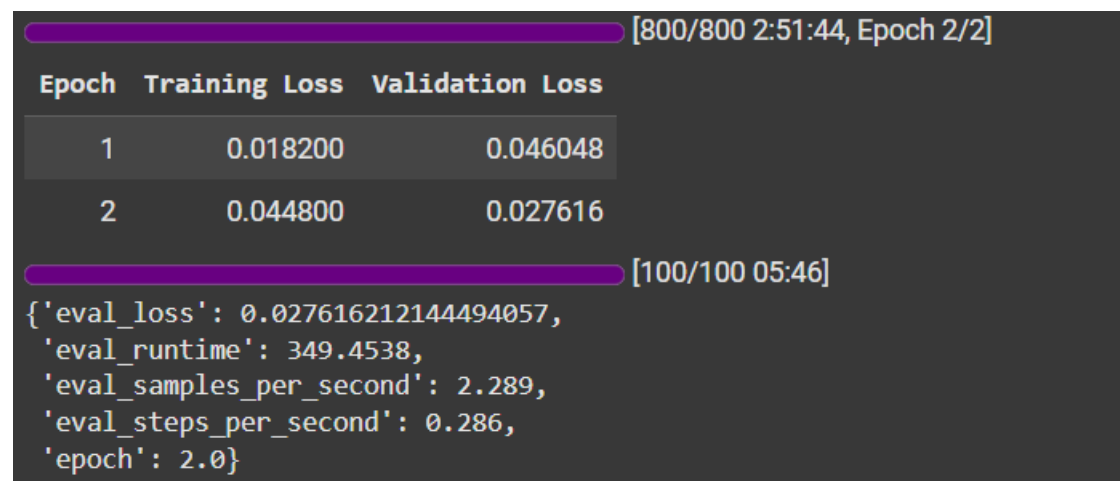
6. washington: 5.440
7. thursday: 5.297
8. friday: 4.993
9. reuters president: 4.794
10. monday: 4.781

Top words predicting FAKE news:

1. video: -10.269
2. image: -8.813
3. via: -8.799
4. gop: -6.093
5. president trump: -6.091
6. hillary: -6.040
7. image via: -5.160
8. obama: -4.930
9. america: -4.494
10. american: -4.481

Deep Learning (BERT) Results

Due to computational constraints, BERT was trained on a subset of the dataset. Despite this limitation, it achieved accuracy and ROC-AUC scores comparable to Logistic Regression, demonstrating its robustness even with reduced data.



A terminal window showing the progress of BERT training. At the top, a purple progress bar is followed by the text "[800/800 2:51:44, Epoch 2/2]". Below this is a table with three columns: "Epoch", "Training Loss", and "Validation Loss". The table contains two rows of data for epochs 1 and 2. Below the table, another purple progress bar is followed by the text "[100/100 05:46]". At the bottom, a JSON object displays various evaluation metrics.

Epoch	Training Loss	Validation Loss
1	0.018200	0.046048
2	0.044800	0.027616

```
{'eval_loss': 0.027616212144494057,  
'eval_runtime': 349.4538,  
'eval_samples_per_second': 2.289,  
'eval_steps_per_second': 0.286,  
'epoch': 2.0}
```

BERT Explainability:

SHAP visualizations for BERT provided interpretability at the token level. For example, in one real news article, terms like “*Trump*” and “*Reuters*” positively influenced a real classification, while in a fake article, words like “*Orlando*” and “*rifle*” contributed to a fake prediction.

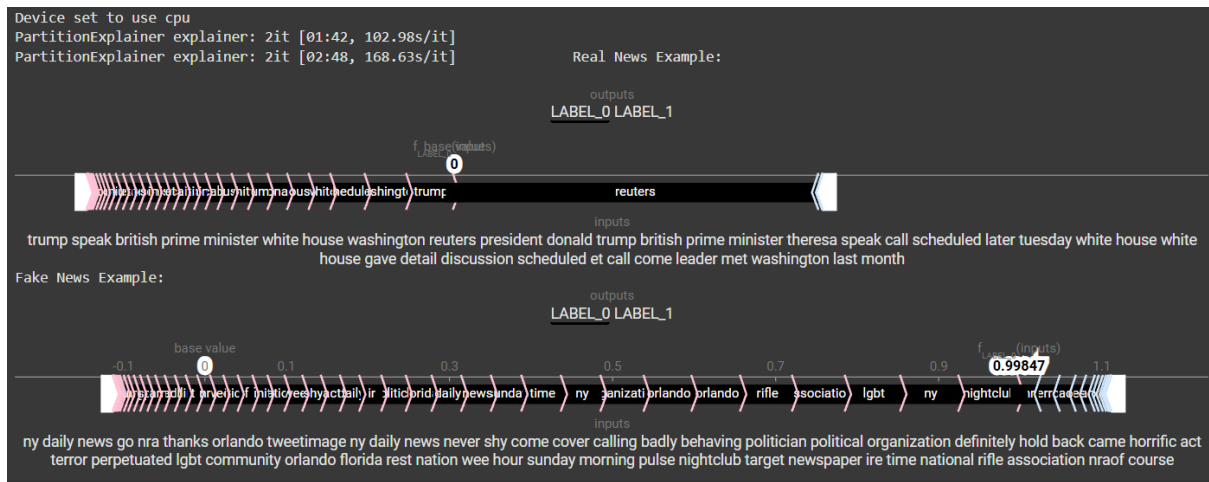


Figure 6. SHAP explanation for one BERT prediction (screenshot/token plot).

Conclusion

This project demonstrates that both Logistic Regression and BERT can effectively detect fake news with high accuracy. Logistic Regression offers a fast, interpretable solution, making it ideal for environments where transparency is critical. BERT, while more computationally demanding, provides cutting-edge performance and is better suited for more nuanced or evolving linguistic contexts. SHAP explainability builds trust by clarifying model decisions.

Assumptions

- The dataset labels are accurate and representative of broader real-world fake news.
- Preprocessing steps retain the essential semantic meaning of the articles.
- Evaluation metrics accurately reflect performance in real-world deployment scenarios.

Limitations

- Potential bias in the dataset may influence model predictions.
- Real-world fake news can differ significantly from the dataset's examples.
- BERT's 512-token limit can truncate long articles, potentially omitting relevant context.
- SHAP computation for deep learning models can be slow, limiting its real-time applicability.

Challenges

- Generalization to new types of fake news.
- Computational cost for BERT and SHAP.
- Explaining decisions to non-technical stakeholders.

Future Uses / Additional Applications

- Browser extension, content moderation, or social media plugin.
- Cross-lingual and multimodal (image/video/text) fake news detection.
- Real-time flagging and user feedback systems.

Recommendations

To maintain performance over time and adapt to evolving news content, I recommend:

1. Regular retraining with fresh data to prevent model drift.
2. Bias monitoring to ensure fairness in predictions.
3. Human-AI collaboration for high-impact decisions, where false positives or false negatives could have serious consequences.
4. Flexible deployment options such as browser extensions, API services, or integration into content moderation systems.

Implementation Plan

1. **Data Cleaning** – Preprocess incoming data for consistency.
2. **Baseline Modeling** – Build a classical model for fast deployment and interpretability.
3. **Deep Learning Enhancement** – Introduce BERT for improved performance where compute allows.
4. **Evaluation and Explainability** – Use ROC-AUC, confusion matrices, and SHAP to ensure transparency.
5. **Deployment** – Offer as a browser extension, API, or moderation tool.
6. **Continuous Improvement** – Retrain and monitor to maintain performance.

Ethical Assessment

- Address potential bias to avoid unfairly flagging certain sources or topics.
- Preserve user privacy by avoiding collection of personal data.
- Maintain transparency by using explainability tools to justify decisions.
- Use human oversight for critical cases.
- Disclose limitations to avoid overreliance on automated systems.

APA References

Bisaillon, C. (2019). *Fake and Real News Dataset* [Data set]. Kaggle.

<https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. <https://doi.org/10.48550/arXiv.1810.04805>

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. <https://doi.org/10.18653/v1/P17-2067>

Zhang, Z., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025. <https://doi.org/10.1016/j.ipm.2019.102025>
